# CART를 이용한 Tree Model의 성능평가

정 용 규* · 권 나 연** · 이 영 호***

목 차



## 요약

데이터 분석가에게 많은 노력이 요구되지 않으면서 사용자가 쉽게 분석결과를 이해할 수 있는 범용 분류기법으로서 가장 대표적인 것은 Breiman이 개발한 의사결정나무를 들 수 있다. 의사결정나무에서 기본이 되는 2가지 핵심내용은 독립변수의 차원 공간을 반복적으로 분할하는 것과 평가용 데이터를 사용하여 가지치기를 하는 것이다. 분류문제에서 반응변수는 범주형 변수여야 한다. 반복적 분할은 변수 의 차원 공간을 겹치지 않는 다차원 직사각형으로 나눈다. 여기서 변수는 연속형, 이진 혹은 서열의 척도이다. 본 논문에서는 새로운 사례를 분류함에 있어서 분류의 성능을 평가하기 위해 분류나무의 정확도 정밀도 재현률 등을 실험하고자 한다.

*표제어: 데이터마이닝, KDD, CART, 분류나무, 회귀나무, Decision Tree, 의사결정,*

접수일(2013년 3월 20일), 수정일(1차: 2013년 3월 26일), 게재확정일(2013년 3월 28일)
* 을지대학교 의료IT마케팅학과 교수, ygjung@eulj.ac.kr
** 을지대학교 의료IT마케팅학과, nayeon090408@gmail.com
*** 수원대학교 컴퓨터학과 외래교수, 교신저자, yhleepr@gmail.com

## 1. 서론

데이터마이닝(Data Mining)이란 대규모 데이터에서 가치 있는 정보를 추출하는 것을 말한다. 즉, 의미심장한 경향과 규칙을 발견하기 위해서 대량의 데이터로부터 자동화 혹은 반자동화 도구를 활용해 탐색하고 분석하는 과정이다. 데이터의 형태와 범위가 다양해지고 그 규모가 방대해지는 빅데이터의 등장으로 데이터마이닝의 중요성은 부각되고 있다. 데이터 분석가에게 많은 노력이 요구되지 않으면서 사용자가 쉽게 분석결과를 이해할 수 있는 범용 분류기법으로서 가장 대표적인 것은 Breiman(1984)이 개발한 나무방법론(tree methodology)이 있다. CART (classification and regression tree; 분류와 회귀나무)는 Breiman 등이 이러한 분류절차를 구현하기 위해 개발한 알고리즘이다.

얼마 전 재형저축이 부활하면서 은행들의 재형저축 유치 경쟁이 과열 양상을 보이자 금융당국이 마케팅 활동을 규제하는 등의 속도 조절에 나서고 있다. 또한 선진 금융을 내세우고 국내에 진입한 외국계금융사에 대한 고객 민원이 95,000여건으로 전년 대비 11.9% 증가했다고 발표했다. 이 중 은행·비은행은 4만 3000건으로 전년 대비 7.0% 늘었다. 이러한 금융상품에 대한 민원이 증가하는 데에 적절한 시장 타켓팅이 이루어지지 않았던 것도 하나의 주요한 이유라고 생각한다. 새로운 상품의 개발 등으로 인해 새로운 시장에 맞는 고객을 대상을 선정하여 은행에서는 서비스를 제공하여야 한다. 따라서 은행은 새로운 금융 상품 등의 마케팅 대상을 선정 할 때에 고객의 거래 상황·능력 등을 엄격히 조사할 필요가 생기게 된다.

본 논문에서 사용한 데이터에서 은행고객들의 나이, 직업, 결혼 유무, 교육수준, 신용, 재산, 주택대출, 개인대출 등을 이용하여 은행고객들이 금융상품에 가입할지의 여부를 분류하는 나무를 보여준다. 분류나무모형이 흔히 사용되는 이유는 나무가 매우 큰 경우에도 나무규칙을 이해하기 쉽기 때문이다. 결국, 분류나무에서 기본이 되는 2가지 핵심내용은 독립변수의 차원 공간을 반복적으로 분할하는 것과 평가용 데이터를 사용하여 가지치기를 하는 것이다. 분류나무의 성과와 관련된 또 다른 문제는 좋은 분류모형을 만들기 위해서는 많은 데이터 집합이 필요하다는 것이다. 최근 Breiman and Cutler는 Random forests를 제시하여 이러한 문제를 분류나무로 확장하여 다루었다. 기본 아이디어는 데이터로부터 다수의 분류나무를 생성하여 그 결과를 더 나은 분류기를 얻기 위해 결합하는 것이다. CART 알고리즘은 출력변수가 연속형인 경우에도 사용될 수 있다. 예측을 위한 회귀나무 역시 분류나무와 동일한 방식으로 작동한다. 회귀나무에서는 출력변수 $Y$가 연속형 변수이고, 원리와 절차 모두 분류나무와 동일하다[1].

## 2. 이론적 배경

### 2.1 Decision tree

데이터마이닝에는 여러 가지 기법들이 있는데 그 중 의사결정트리는 데이터마이닝 분석의 대표적인 분석 방법이다. 인공지능, 기계학습, 통계분석에서도 역시 의사결정트리 알고리즘은 활용이 많이 되고 있다. 의사결정트리는 간단하게 결정트리(Decision Tree)라고 불리기도 한다.

과거에 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 패턴, 즉 분류별 특성을 속성의 조합으로 나타내는 분류모형을 나무의 형태로 만드는 것이다. 의사결정 규칙을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류(Classification)하거나 예측(Prediction)을 수행하는 계량적 분석 방법이다.

의사결정트리는 새로운 레코드의 분류 값을 예측하기 위해 이미 만들어진 분류모형(결정나무)이 지시하는 바에 따라 레코드의 속성 값을 질문하는 작업

을 반복적으로 수행한다. 특히 결정적인 질문을 던지게 되면 다른 모든 속성의 값을 묻지 않고도 레코드이 분류 값을 정확히 예측할 수 있다. 따라서 레코드를 분류하고 예측할 수 있는 나무(모형)를 얼마나 잘 만드느냐가 의사결정트리의 핵심이다.

### 2.2 CART

CART(Classification And Regression Tree)는 각각 종속 변수가 범주 또는 숫자인지 여부에 따라 분류 또는 회귀 나무 중 하나를 생성하고 비모수 결정나무 학습법이다. CART 알고리즘은 의사결정나무 분석을 형성하는데 있어서 가장 보편적인 알고리즘이라고 할 수 있다. 1984년 Briemen에 의해 발표되어 Machine-learning 실험의 시초가 되고 있다.

모형의 형성은 training data set을 가지고 한다. 목표변수는 이미 그 분류가 알려져 있으며, 나머지 설명변수를 가지고 이 목표변수를 잘 분류할 수 있는 모형을 만들어 새로운 데이터 세트에 적용시킨다. CART 알고리즘은 이진트리구조로 모형을 형성하는데 목표변수를 가장 잘 분리하는 설명변수와 그 분리시점을 찾는 것이다.

처음 분리기준으로 두 개의 마디를 형성하면, 목표변수를 한쪽의 값으로 분류시킬 수 있는 기준을 찾아서 계속 나무를 형성해 나간다. 그래서 더 이상의 분리가 이루어지지 않고 다양성이 효과적으로 줄었을 때 끝 노드를 형성한다. 이렇게 완전히 Full tree를 형성하는 것은 주어진 데이터는 잘 맞추겠지만, 새로운 데이터가 들어오면 잘 분류하지 못할 수 있다. 새로운 데이터 셋이 들어와도 그 예측을 일반적으로 잘 할 수 있도록 적절한 가지치기를 해주어야 한다.

### 2.3 회귀분석 모형(Regression Model)

데이터마이닝 분석도구로서 가장 먼저 고려되는 것은 예측(Prediction and Regression)과 회귀분석모형이다. 통계학에서 데이터마이닝 도구로 가장 많이 사용되는 모형들 중 하나가 회귀분석(regression)이다. 대표적인 예측 모형으로는 선형회귀(linear regression)과 로지스틱 회귀(logistic regression)가 있다. 데이터마이닝뿐만 아니라 대부분의 데이터 분석 과정에서 우선적으로 산점도(scatter plot)와 같은 데이터의 시각화(visualization)를 통하여 변수들 간의 대략적인 관계를 파악할 수 있다. 선형 회귀분석(linear regression)은 목표변수가 연속형(continuous)인 지도학습 예측모형(supervised prediction)이다. 데이터마이닝 관점에서 선형 회귀분석의 목표는 주어진 입력값에 대한 목표값을 예측할 수 있는 모형을 구축하는 것이다.

### 2.4 Bayes's rule

통계학에서 데이터를 분석하는 하나의 방법론이라고 할 수 있다. 데이터를 분석할 때 관측된 데이터만 가지고 분석을 하게 되면 정확도가 떨어질 뿐만 아니라 여러 한계점이 드러나게 된다. 이에 반면 Bayes's rule을 기반으로 데이터를 분석하면 과거에 이미 알려진 사전확률을 고려함과 동시에 각 상황에 따라 분석자의 주관적인 생각까지 함께 분석을 하게 되므로 기존의 분석방법에 비해 훨씬 더 정확한 결론을 얻을 수 있게 된다. 독립적인 가정이 분명하게 위반되었을지라도 Naive Bayes는 놀라운 효과적인 작용을 나타낸다. 왜냐하면 분류는 정확한 확률을 산정해내는 것을 필요로 하지 않으며 최대확률은 올바른 클래스로 지정된다. 그러나 동일한 속성과 같은 필요이상으로 많은 중복적인 속성들은 문제를 일으키며 또한 많은 수치적인 속성들은 평균적으로 분포될 수가 없다.

### 2.5 분류나무

분류나무란 의사결정규칙(decision rule)을 나무구조로 도표화하여 분류(classification)와 예측(prediction)을

수행하는 분석방법이다. 분류나무의 핵심적인 두 가지는 독립변수의 차원을 반복적으로 분할하는 것과, 평가용 데이터를 사용하여 가지치기를 하는 것이다.

CART 알고리즘은 출력변수가 연속형인 경우에도 사용될 수 있다. 과정은 다음과 같다. 회귀나무에서는 출력변수 Y가 연속형 변수이고, 원리와 절차 모두 분류나무와 동일하다. 많은 분할이 이루어지고, 각 분류별로 나무의 각 가지에서의 불순도를 측정한다. 그리고 불순도의 합이 최소가 되는 분할을 선택한다. 회귀나무와 분류나무는 예측, 불순도 측정, 성과평가 등 3가지 측면에서 차이가 있다. 예측은 분류나무의 경우 끝마디에 속한 학습용 데이터 중에서 집단의 수가 가장 큰 집단으로 '다수결'에 의해 결정된다. 회귀나무에서는 끝마디의 값은 끝마디의 학습용 데이터의 평균값으로 결정된다. 분류나무의 불순도 측정치로는 지니계수, 엔트로피 지수가 있다. 두 지수는 모두 해당 마디의 관찰치들의 범주 사이의 비율함수로 정의된다. 회귀나무에서 학습용 집합의 각 가격으로부터 집단평균값의 편차제곱을 모두 더한 값이다. 마디의 모든 값들이 같을 때, 불순도는 가장 작은 값인 0을 갖는다. 앞에서 언급했듯이 예측은 마디의 출력변수 값의 평균을 통해 얻어진다. 본 논문에서는 일반적인 예측과 오차에 대한 정의로서 회귀나무의 예측성과는 다른 예측기법과 마찬가지로 평균제곱오차의 제곱근(root-mean-squared error)과 같은 요약측정치와 리프트 도표(lift chart)와 같은 그래프를 이용하여 평가한다[2].

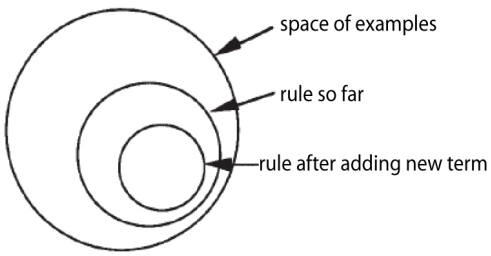


그림 1. 커버링알고리즘에서 개체공간
Fig. 1. Instance Space of Covering Algorithm

## 3. 연구모형

본 논문의 실험은 속성값 중 marital, housing, loan, duration, poutcome을 사용하여 클래스 변수인 y, 즉 계좌의 여부를 확인한다. 데이터에는 age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome, y로 구성되어 총 17개의 속성값으로 45,211개의 인스턴스로 구성되어 있는 데이터이다. 실험을 위해서 속성 선택 필터를 통한 전처리 과정을 거쳤다.

## 4. 실험 및 결과분석

실험은 weka를 사용하였으며, 실습 데이터는 Bank Marketing을 사용하였다. 45211개의 학습데이터를 사용하여 실험을 진행하였다.

표 1. 데이터 각 attribute 속성
Tab. 1. Each Data Attribute Properties

| | |
|---|---|
| Age | Customer's Age in completed years |
| Job | Customer's job category |
| marital | Customer's marital status (married, divorced, single) |
| education | education category (unknown, secondary, primary, tertiary) |
| default | has credit in default? (yes, no) |
| balance | average yearly balance, in euros |
| housing | has housing loan? (yes, no) |
| loan | has personal loan? (yes, no) |
| contact | contact communication type category (unknown, telephone, cellular) |
| day | last contact day of the month |
| month | last contact month of year |
| duration | last contact duration, in seconds |
| campaign | number of contacts performed during this campaign for this client |
| pdays | number of days that passed by after the client was last contacted from a previous campaign |
| previous | number of contacts performed before this campaign and for this client |
| poutcome | outcome of the previous marketing campaign category(unknown, other, failure, success) |
| y | has the client subscribed a term deposit? (yes, no) |

표 1은 Bank Marketing의 각 attribute 속성값의 의미를 설명하고 있다. 본 논문의 실험에서는 위 속성들 중 상품 수락 여부에 불필요한 속성들을 제거 후 실습하였다.

표 2. 전처리 후 데이터의 attribute 속성

Tab. 2. After Prepocessing Data Attribute Properties

| | |
|---|---|
| marital | Customer's marital status (married, divorced, single) |
| housing | has housing loan? (yes, no) |
| loan | has personal loan? (yes,no) |
| duration | last contact duration, in seconds |
| poutcome | outcome of the previous marketing campaign category(unknown, other, failure, success) |
| y | has the client subscribed a term deposit? (yes, no) |

표 3. 실험 결과 요약

Tab. 3. Summary

| | | |
|---|---|---|
| Correctly Classified Instances | 40728 | 90.0843% |
| Incorrectly classified Instances | 4483 | 9.9157% |
| Kappa statistic | 0.4005 | |
| Mean absolute error | 0.1575 | |
| Root mean squared error | 0.281 | |
| Relative absolute error | 76.2155% | |
| Root relative squared error | 87.4359% | |
| Total Number of Instances | 45211 | |

본 실험에서 attribute selection 및 실험의 정확성을 증가시키기 위한 전처리 과정으로 10-fold cross-validation을 조건으로 지정하여 실험을 진행하였다. 위의 표 3은 실험결과의 요약된 결과이다. 총 45,211개의 instance로 실험한 결과 y에 대응되도록 일치된 개체는 40,728개로 약 90.1%의 정확성을 보이고 있다.

표 4. 혼동 행렬

Tab. 4. Confusion Matrix

| a | b <-- classified as | |
|---|---|---|
| 38896 | 1026 | a(= no) |
| 3457 | 1832 | b(= yes) |

아래 그림은 회귀나무를 그리기 위한 Full Tree Rules이다. 그림에서 Level은 Tree의 높이고, SplitVar는 트리를 분류하는 속성이다. SplitValue는 SplitVar를 기준으로 분류하는 속성들의 값이며, Cases는 분류되는 데이터의 수이다. Node Type을 보면 Decision으로 분류된 것을 확인 할 수 있다. 위의 분류 기준값으로 회귀나무를 그려보면 다음과 같은 결과를 확인 할 수 있다.

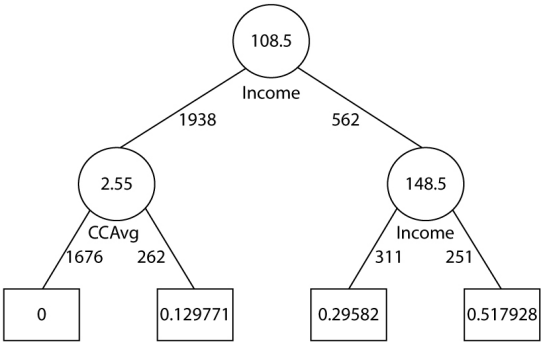


그림 2. 회귀나무 결과

Fig. 2. Result of Regression Tree

위 그림의 회귀나무 결과를 보면 레벨 1에서 수입(Income) 108.5를 기준으로 대출 2500개의 학습데이터 중 1938개와 562개의 값을 나누는 것을 확인할 수 있다. 레벨 2에서는 카드평균사용 2.55를 기준으로 1676개의 데이터가 대출을 받지 않았다는 것을 알 수 있고, 262개가 0.129771의 확률로 대출을 받았다는 것을 알 수 있다. 또한 오른쪽에서 수입 148.5를 기준으로 311명은 0.29582의 확률로 대출을 받고 251명은 0.517928의 확률로 대출을 받은 것을 알 수 있다.

## 5. 결론

실험 결과 분류와 회귀나무는 결측치를 대체하거나 결측값을 가진 관찰치를 삭제하지 않아도 결측데이터를 처리할 수 있다는 것을 확인할 수 있었다.

따라서 이 방법은 해당 변수가 분류성능에 미치는 영향도의 관점에서 변수들의 중요도 순위를 평가하는 목적으로 확장될 수 있다. 계산측면에서는 나무모형은 모든 변수에 대해 모든 가능한 분리를 계산하기 위해서 수많은 정렬계산과정을 필요로 하기 때문에 학습하는데 상대적으로 많은 시간이 소요된다. 또한 가지치기를 할 경우 더 많은 계산시간이 소요될 수 있다. 그러나 회귀나무의 가장 실제적인 이점은 규칙을 이해하기 쉽게 표현하는 것이라고 할 수 있다.

## 참 고 문 헌

〔국외 문헌〕

[1] Galit Shmueli, Nitin R. Patel, and Peter C. Bruce (2009), "Data Mining for Business Intelligence", Wiley, 2009.

[2] Pang-Ning Tan (2007), Michael Steinbach, Vipin Kumar, "Data Mining", Addison Wesley.

[3] Jesse Davis, Vítor Santos Costa, Irene M. Ong, David Page and Inês Dutra (2004), "Using Bayesian Classiers to Combine Rules", Department of Biostatistics and Medical Informatics, University of Madison-Wisconsin.

[4] Jesús Cerquides (2003), "Tractable Bayesian Learning of Tree Augmented Naive Bayes Classifiers", Ramon López de Màntaras.

[5] Stuart Moran, Yulan Hey, Kecheng Liu (2009), "An Empirical Framework for Automatically Selecting the Best Bayesian Classifier", Proceedings of the World Congress on Engineering, Vol. I, WCE 2009, July 1-3.

**정 용 규 (Yong Gyu Jung)**

서울대학교, 연세대학교, 경기대학교에서 각각 학사, 석사, 박사학위를 취득하였고, 현재 을지대학교 의료IT마케팅학과 교수로 재직 중이다. ISO 및 UN에서 전자거래분야 한국대표위원으로 활동하고 있으며, 의료정보, 전자무역, 해상물류, 금융전산에 Semantic Web, Process Modelling, ebXML 등의 표준기술의 적용에 관심이 많다.

**권 나 연 (Na Yeon Kwon)**

을지대학교 의료IT마케팅학과에 재학 중이며, 의사결정트리 ID3와 C4.5 알고리즘의 성능비교 등을 연구하고 있다. 병원에서 사용하고 있는 OCS, EMR 등 의료정보시스템 구현기술과 임상데이터 분석을 위한 데이터 마이닝 기술에 관심이 많다.

**이 영 호 (Young Ho Lee)**

수원대학교에서 학사, 석사, 박사를 취득 및 수료하였다. 현재 스마트주소 위원회(COCif)에서 한글주소체계 기반 다이얼 서비스 분야 국가표준을 개발하고 있다. 주요 관심분야로는 분산 네트워크, RFID 기반의 상품 관리 및 검색, 휴대전화를 위한 하이브리드형 한글 입력 방식 및 키패드 배열 설계 등에 관심이 많다.

# Using CART to Evaluate Performance of Tree Model

Yong Gyu Jung* · Na Yeon Kwon** · Young Ho Lee***

## ABSTRACT

Data analysis is the universal classification techniques, which requires a lot of effort. It can be easily analyzed to understand the results. Decision tree which is developed by Breiman can be the most representative methods. There are two core contents in decision tree. One of the core content is to divide dimensional space of the independent variables repeatedly, Another is pruning using the data for evaluation. In classification problem, the response variables are categorical variables. It should be repeatedly splitting the dimension of the variable space into a multidimensional rectangular non overlapping share. Where the continuous variables, binary, or a scale of sequences, etc. varies. In this paper, we obtain the coefficients of precision, reproducibility and accuracy of the classification tree to classify and evaluate the performance of the new cases, and through experiments to evaluate.

*Keywords: Datamining, KDD, CART, classification trees, regression trees, Decision Tree, Decision-making*

* Eulji University, Dept. of Medical IT Marketing, ygjung@eulj.ac.kr
** Eulji University, Dept. of Medical IT Marketing, nayeon090408@gmail.com
*** Suwon University, Dept. of Computer Science, Corresponding author, yhleepr@gmail.com